Making the Critical Appraisal for Summaries of Evidence (CASE) for evidence-based medicine (EBM): critical appraisal of summaries of evidence*

Margaret J. Foster, MLIS, MPH, AHIP; Suzanne Shurtz, MLIS, AHIP

See end of article for authors' affiliations.

DOI: http://dx.doi.org/10.3163/1536-5050.101.3.008

Objectives: Standards for evaluating evidence-based medicine (EBM) point-of-care (POC) summaries of research are lacking. The authors developed a ''Critical Appraisal for Summaries of Evidence'' (CASE) worksheet to help assess the evidence in these tools. The authors then evaluated the reliability of the worksheet.

Methods: The CASE worksheet was developed with 10 questions covering specificity, authorship, reviewers, methods, grading, clarity, citations, currency, bias, and relevancy. Two reviewers independently assessed a random selection of 384 EBM POC summaries using the worksheet. The responses of the raters were then compared using a kappa score.

Results: The kappa statistic demonstrated an overall moderate agreement ($\kappa=0.44$) between the reviewers using the CASE worksheet for the 384 summaries. The 3 categories of evaluation questions in which the reviewers disagreed most often were citations ($\kappa= 0$), bias ($\kappa=0.11$), and currency ($\kappa=-0.18$).

Conclusions: The CASE worksheet provided an effective checklist for critically analyzing a treatment summary. While the reviewers agreed on worksheet responses for most questions, variation occurred in how the raters navigated the tool and interpreted some of the questions. Further validation of the form by other groups of users should be investigated.

INTRODUCTION

Evidence-based practice ''integrates the best external evidence with individual clinical expertise and patients' choice'' [1]. When assessing the evidence, readers can use critical appraisal tools (CATs) to guide them in determining the reliability and validity of studies. Several organizations have provided critical appraisal sheets online, such as the popular checklists from the Critical Appraisal Skills Programme (CASP) international network [2]. Questions on the appraisal sheets, also called crib sheets, vary depending on the study type (systematic review, randomized control trial, etc.). These worksheets have become common tools to train health sciences students in how to review research articles [3–5]. Practicing health care providers also utilize critical appraisal worksheets to assess the quality of evidence [6, 7].

Highlights

- Few critical appraisal tools have been evaluated with inter-rater reliability testing.
- The ways that users of evidence-based medicine (EBM) point-of-care (POC) tools interpret how to appraise an evidence summary—particularly when defining the grading of evidence, currency, and bias—may vary even when a standard evaluation sheet is used.
- The Critical Appraisal for Summaries of Evidence (CASE) worksheet had a moderate level of inter-rater reliability, similar to previous evaluative studies of critical appraisals tools.

Implications

- Medical librarians can develop tools useful for librarians, students, and clinicians to guide them in appraising clinical evidence summaries.
- The CASE worksheet can be a valuable tool to consider the quality of individual evidence summaries and to see patterns of overall quality in EBM POC tools.

While there are worksheets for appraising many types of evidence, currently no critical appraisal sheet exists for evidence summaries in evidence-based medicine (EBM) point-of-care (POC) tools. These resources are often consulted in clinical settings to assist in making patient care decisions. Rather than searching and reviewing the literature themselves, busy clinicians utilize such tools as DynaMed, UpToDate, and Epocrates to find summaries of research and recommendations based on that evidence. In an earlier study, the authors completed an evaluation of EBM POC tools [8]. During this evaluation process, it became evident that, while more and more POC tools are being developed, there are no standards set for how these tools gather or grade evidence. In addition, the researchers observed that, even within a single POC tool, the summaries varied widely in the

* Based on a presentation at SCC MLA 2011, Annual Meeting of the South Central Chapter (SCC) of the Medical Library Association; Baton Rouge, LA; 18 October 2011.

A supplemental appendix is available with the online version of this journal.

Table 1
Systematic reviews on evaluation of critical appraisal tools (CATS)

Author	Resources searched	CATs included	Evaluation of tools	Number evaluating reliability	Number providing overall inter-rater reliability
Crowe & Sheppard (2011) [9]	CSA Illumina, EBSCOhost, Gale InfoTrac, Informit, ISI Web of Knowledge, JSTOR, OvidSP, ProQuest, Scopus, Cochrane Library*	44 critical appraisal tools	content analysis, descriptive analysis of structure, methods, analysis of data used	10	6
Graham et al. (2000) [10]	MEDLINE, reference lists, contacted authors	13 clinical practice guideline appraisal instruments	content analysis, audience of tool, purpose, development, validation	4	0
Katrak et al. (2004) [11]	TRIP, Clinical Evidence, Physiotherapy Evidence Database, OT Seeker, McMaster University Evidence-Based Practice Group, MEDLINE, EMBASE, CINAHL, Current Contents, Cochrane Library, DARE, SIGN, National Health and Medical Research Council (NH&MRC), Google, Yahoo, MSN, reference list, contacted authors	121 tools on various designs for allied health	content analysis, method of development, method of evaluating validity and quality	11	4
Vlayen et al. (2005) [12]	MEDLINE, EMBASE, CINAHL, personal contacts	24 tools for clinical practice guidelines	content analysis	4	1
Wendt & Miller (2012) [13]	CINAHL, ERIC, Linguistics & Language Behavior Abstracts (LLBA), MEDLINE, PsycINFO, Google Scholar, Ixquick, ScienceDirect, Scirus, Scopus, SpringerLink	7 critical appraisal tools for single-subject experimental designs	content analysis, validity, reliability	3	1
Total				28*	7*

* Minus 5 duplicates.

type of information included [8]. Just as critical appraisal worksheets have been developed to guide readers in evaluating the quality of studies, the researchers recognized the need for a guide to assess the quality of the evidence in POC tools. The authors sought to develop this guide, which they named the ''Critical Appraisal of Summaries of Evidence'' (CASE) worksheet, to be utilized by anyone needing to evaluate a treatment summary provided in POC tools, including health care professionals, students, and librarians.

Research question

The main research question for this project is: Can a reliable tool be designed to critically analyze treatment summaries of evidence-based POC tools? This question is addressed through two objectives:

- to develop a form to critically analyze treatment summaries in POC tools
- to evaluate the inter-rater reliability of the form

Literature review

The authors began by seeking best practices in creating and evaluating the inter-rater reliability of critical appraisal tools. A search was conducted in MEDLINE, EMBASE, CINAHL, PsycINFO, and ERIC yielding 5 systematic reviews and over 100 primary studies. The primary studies are discussed in the discussion section as a comparison to results of this study. The methods and findings of the systematic reviews are summarized in Table 1. As illustrated in the table, only a small percentage of CATs in the systematic reviews described the tool development process (35 out of 209) or evaluated reliability (32 out of 209) [9–13]. The largest of these reviews, Katrak et al., evaluated 121 articles on critical appraisal sheets and observed a great variety in these sheets partly due to a lack of a gold standard [11]. Crowe and Sheppard closely studied the development of 44 CATs and stated that ''CATs are being developed while ignoring basic research techniques [or] the evidence available for design'' [9]. Content analysis was conducted on some level by all of the reviews [9–13]. Vlayen et al., who reviewed articles appraising clinical guideline tools, concluded that ''to evaluate the quality of the clinical content and more specifically the evidence base of a clinical practice guideline, verification of the completeness of the quality of the literature search and its analysis has to be added to the process of validation by an appraisal instrument'' [12].

The present authors acknowledged the importance of including questions on the CASE worksheet that address how evidence is collected and evaluated in POC tools. All of the reviews noted that few CATs had been evaluated for reliability [9–13], with Wendt and Miller suggesting that ''researchers and practitioners are well advised to proceed with caution when selecting quality appraisal tools for EBP needs'' and should be reviewing the validity and reliability of the instrument [13]. Graham et al. went further to suggest that CATs should be compared in experimental studies to determine the best CATs for different uses

Figure 1
Critical Appraisal for Summaries of Evidence (CASE) worksheet

Questions	Evaluation
Summary topic	
1. Is the summary specific in scope and application?	Yes Not completely No
Summary methods	
2. Is the authorship of the summary transparent?	Yes Not completely No
3. Are the reviewer(s)/editor(s) of the summary transparent?	Yes Not completely No
4. Are the search methods transparent and comprehensive?	Yes Not completely No
5. Is the evidence grading system transparent and translatable?	Yes Not completely No
Summary content	
6. Are the recommendations clear?	Yes Not completely No
7. Are the recommendations appropriately cited?	Yes Not completely No
8. Are the recommendations current?	Yes Not completely No
9. Is the summary unbiased?	Yes Not completely No
Summary application	
10. Can this summary be applied to your patient(s)?	Yes Not completely No

[10]. The present researchers concluded that the CASE worksheet should be developed using the best evidence available and that it was essential to measure the inter-rater reliability to demonstrate the quality of the tool.

METHODS

Development of the Critical Appraisal for Summaries of Evidence (CASE) worksheet

The next step was to develop the CASE worksheet. Using insights gained from previous research in evaluating POC tools and reviewing existing critical appraisal sheets, the authors brainstormed what criteria would establish the quality of an evidence summary. They then wrote ten questions evaluating these criteria, keeping in mind the kinds of questions most helpful for those who use POC tools to make clinical decisions. The questions were then classified under four headings: summary topic, summary methods, summary content, and summary application. The included questions could be answered with a ''yes,'' ''no,'' or ''not completely'' (where the information was incomplete) (Figure 1). Under each main question, guiding questions were provided to assist the evaluator in thinking through the answer. These guiding questions are provided in the final version of the CASE form (Appendix, online only).

Sample

The next step was to determine which POC tools would be used to test the inter-rater reliability of the CASE worksheet and how to find a random sample of evidence summaries from these tools. The investigators opted to include all POC tools that had summaries of treatment and claimed to be evidence based. To gather this list, the researchers searched for names of POC tools in journal articles and on the Internet, and looked at the tools that other health sciences libraries subscribed to. POC tools selected for the evaluation were: 5 Minute Clinical Consult, ACP Pier, DynaMed, eMedicine, Epocrates, Essential Evidence Plus, First Consult, and UpToDate. Originally, Clinical Knowledge Summaries from the National Health Service and Harrison's Practice product, ''Answers on Demand,'' were included in the evaluation list, but over the course of the project, these tools were no longer being updated or became unavailable, so they were ultimately excluded from the final list. The researchers signed up for trials to have access to the tools that were not freely available and to which their institution did not subscribe.

Once the tools were selected, the authors then determined how to best sample the treatment summaries within the tools. To determine the size of a statistically significant sample, the total number of treatment summaries available in each tool was gathered by viewing marketing information, usually on the tools' websites. After adding the total estimated number of summaries available (27,695), a sample size was calculated (379) that achieved a confidence interval of 95%. Due to rounding, the final sample size was set at 384 summaries (Table 2). The number of individual summaries to be evaluated in each tool was proportional to the total summaries of all the tools.

Summaries were randomly selected from the tools, with the method of randomization determined after reviewing the structure of the product. Many of the POC tools provided browsing capabilities with only two labeling each summary with a record number. Most browsing was available alphabetically and sometimes categorically. If record numbers were available, these were employed in the randomization.

Table 2
Number needed to evaluate from each evidence-based medicine (EBM) tool

EBM tool	No. of estimated summaries	% of all summaries	Calculated
5 Minute Clinical Consult	900	3%	12
ACP PIER	368	1%	6*
DynaMed	3,200	12%	44
eMedicine (Medscape)	8,293	30%	114*
Epocrates	1,667	6%	24*
Essential Evidence Plus	2,735	10%	38*
First Consult	1,532	6%	22*
UpToDate	9,000	32%	124*
Total	27,695	100%	384

* Rounded up 1.

Table 3
Number of times answers agreed on and kappa for questions 1–9 (n=384 forms)

Question	Yes		Not complete		No		Kappa	(Interpretation)
Specificity	382	(99%)	0	(—)	0	(—)	0	(paradox)*
Authorship	286	(74%)	0	(—)	68	(18%)	0.80	(good)
Reviewers	264	(69%)	9	(2%)	45	(12%)	0.62	(good)
Search methods	19	(5%)	0	(—)	336	(88%)	0.54	(moderate)
Grading	94	(24%)	3	(1%)	157	(41%)	0.38	(fair)
Clarity	324	(84%)	0	(—)	0	(—)	0	(paradox)*
Citations	160	(42%)	10	(3%)	9	(2%)	0	(poor)
Currency	162	(42%)	0	(—)	34	(9%)	0.11	(poor)
Bias	41	(11%)	12	(3%)	7	(2%)	−0.16	(poor)
Overall	1,732	(50%)	34	(1%)	651	(19%)	0.44	(moderate)

* Agreement was high, but kappa was low.

Otherwise, browseable lists were copied into MS Excel, and each link was given a record number to be randomized.

Statistics and software

The CASE worksheet was created as a form in Zoho Creator, a free online service for creating databases of forms. An MS Excel spreadsheet was used to track which summaries were to be completed for which tools. Two reviewers independently completed the CASE worksheet form for each of the 384 randomly selected treatment summaries, for a total of 768 appraisals. Question 10 of the CASE worksheet, application of the summary to the patient, was omitted as there was no specific patient in this scenario. Once the raters evaluated the summaries using the CASE form, the data were exported from Zoho into MS Excel, and the Fleiss-kappa (κ) score measuring the inter-rater reliability for each question was calculated.

RESULTS

The 2 evaluators each completed the CASE form assessing 384 randomly selected summaries, for a total of 768 summaries. The κ calculated overall was $\kappa=0.44$, which is interpreted as fair agreement (Table 3). The questions that the evaluators agreed on most frequently, receiving a κ score of good agreement, were on authorship ($\kappa=0.80$) and reviewers ($\kappa=0.62$). The questions on which the evaluators agreed a fair amount of time were regarding search methods ($\kappa=0.54$) and grading of evidence ($\kappa=0.38$). A paradox occurred for questions 1 (specificity) and 6 (clarity). This paradox has been reported in several other studies [14, 15]. The paradox occurs when answers are not symmetrically distributed across categories, causing κ to be very low, while agreement level was actually high. For instance, with question 6 of this study, both authors responded ''yes'' 324 times to the question (84%); however, the kappa was calculated as 0; thus, high agreement, but low kappa. The questions on which evaluators disagreed most frequently regarded citations ($\kappa=0.05$), currency ($\kappa=0.11$), and bias ($\kappa=-0.18$). Table 4 shows the number of times the evaluators agreed on each question.

Table 4 shows the percentage, by POC tool, with which the evaluators answered ''yes'' to the CASE questions. The category of specificity had the highest percentage (100%) of ''yes'' answers for all summaries viewed in all tools. The next highest average (90%) was the question on clarity. Other categories of questions averaged from 35% to 65% in terms of ''yes'' responses across tools. The lowest question category to receive a ''yes'' response was for search methods (9% average). The question of authorship received a total average of 63% of ''yes'' responses, but with the widest range of averages (0 to 100%), with ACP Pier, eMedicine, Epocrates, First Consult, and UpTodate being the tools that consistently received ''yes'' responses in this category.

Table 4
Percentage of summaries in EBM tool with ''yes'' for each question

Questions	5 Minute Clinical Consult	ACP Pier	DynaMed	eMedicine	Epocrates	Essential Evidence Plus	First Consult	UpToDate	Average for question
# of summaries analyzed	24	12	88	228	48	76	44	248	
1. Specificity	100%	100%	100%	100%	100%	100%	100%	100%	100%
2. Authorship	4%	100%	1%	100%	100%	—	93%	100%	62%
3. Reviewers	—	92%	15%	99%	98%	—	—	100%	50%
4. Search methods	—	—	63%	—	2%	5%	—	—	9%
5. Grading	25%	75%	73%	—	29%	100%	34%	48%	48%
6. Clarity	79%	100%	83%	95%	98%	86%	86%	96%	90%
7. Citations	25%	100%	55%	50%	85%	95%	41%	70%	65%
8. Currency	17%	67%	77%	73%	75%	1%	45%	72%	53%
9. Bias	29%	100%	19%	—	4%	50%	32%	48%	35%

Table 5
Primary studies on overall inter-rater reliability of CATs measured by kappa score

Author (year)	CAT assessed	# of raters	Items assessed	# of assessments	Overall inter-rater reliability
Hollingworth et al. (2006) [16]	Quality Assessment of Diagnostic Accuracy Studies (QUADAS)	6	19	38	κ=0.22
Hoy et al. (2012) [17]	Modified version of Leboeuf-Yde and Lauritsen tool	2	54	108	κ=0.82
Kang et al. (2012) [18]	AMSTAR	2	41	84	κ=0.50
Shaneyfelt et al. (2006) [19]	Assessment tool of clinical guidelines	2	279	558	κ=0.73
Average kappa					0.57

DISCUSSION

Findings

The main finding of the inter-rate reliability calculation is that the CASE form has a moderate level (κ=0.44) of agreement between 2 experienced medical librarians who rated 384 treatment summaries in 8 different POC tools for the first 9 questions. In comparing the overall κ of the present study with similar research (Table 5), the average κ for the 4 studies found was 0.57, which was also in the same range of moderate agreement [16–19]. As this was the researchers' first attempt to measure the κ of the CASE worksheet, this was an indicator that with minor revisions, a higher level of reliability for the tool is achievable.

The κ scores elucidated questions in which responses varied widely. A discussion between the evaluators after reviewing the results clarified why they might have chosen differently for certain questions. For questions achieving fair to moderate agreement, such as search methods and grading, dialogue after the review clarified that the evaluators might have looked in different places in the tools to find answers to these questions. As for differences in questions relating to whether or not the evidence was specifically graded, a couple of tools took some drilling down to locate a grade, which might have been missed by one of the reviewers.

The researchers also realized that, although they had coauthored the questions, when trying to apply the questions to specific treatment summaries, each interpreted the questions and the response options differently. For example, the evaluators varied in how they responded ''Yes,'' ''Not completely,'' and ''No'' when looking at how much evidence was assigned a grade or when seeing how many author or reviewer conflicts of interest were listed. The raters varied in how they quantified the amount of graded evidence when determining a ''Yes'' score versus a ''Not completely'' score. The evaluators also discovered they had applied different standards in judging currency. One evaluator gave a summary a ''Yes'' for currency to summaries updated within the last 6 months, a ''Not completely'' response to those updated within the last year, and a ''No'' response to summaries updated more than a year ago. The other evaluator chose a ''Yes'' response for summaries updated within the last year, ''Not completely'' for summaries updated within the last two years, and a ''No'' to summaries updated further back than two years.

From the limited summaries evaluated, patterns in responses to questions about the tools emerged. The evaluators found that authorship of summaries tended to be either completely transparent or not at all. Because of this, it was often difficult to determine bias of evidence summaries. Also striking in the findings was the observation that DynaMed was the only tool for which the evaluators consistently found search methods (Table 4).

The researchers concluded that if they, as creators of the CASE worksheet, interpreted some of the questions and answers differently from each other, then potential users would also vary in how they completed the worksheet. For instance, the definition of what is ''current'' evidence may differ widely from user to user. To retain the original intention of the questions and to increase the reliability of the worksheet, the researchers revised the questions where the greatest variance had occurred in the evaluation.

Revisions to CASE worksheet

Based on the reliability study, revisions were made to the questions receiving ''poor'' agreement between evaluators. These questions related to the grading of the evidence, currency of the summary, and possible bias or conflicts of interest of summary authors or reviewers. The researchers first assessed how to clarify the CASE worksheet questions on grading. To remove ambiguity, the main question was changed to ''Is the evidence graded and is the system transparent and translatable?'' (Appendix, online only). The researchers felt that this question led to a more clear-cut ''Yes'' or ''No'' answer. They also revised one of the guiding questions to be more direct, ''Is there a grade for each recommendation and/or cited study?'' The goal of this revision was to remove doubts as to how to quantify how much of a summary is graded to earn a ''Yes'' response.

For the CASE worksheet section on summary currency, researchers first removed the ''Not completely'' response. Originally, they had wanted to provide consistent options throughout all sections of the worksheet, but upon evaluation, a ''Not completely'' response led to confusion in terms of currency. After discussion as to a standard for ''current'' evidence, the researchers elected to follow the bar set in the *Cochrane Handbook for Systematic*

Reviews of Interventions, which requires that evidence be updated every two years [20]. Once this consensus was reached, the CASE worksheet section on currency was edited accordingly. The guiding questions were deleted and replaced with, ''Has the summary been updated within the last 2 years?''

Finally, the researchers revised the section on bias to also be more clear-cut. The researchers became convinced that the two guiding questions were what had led to the variance in responses: ''Is there a conflict of interest between the recommendations of the summary and the sponsor? Are conflicts of interest of summary author(s) and reviewer(s) provided?'' After some discussion, the evaluators realized that the purpose of these questions was for users to notice if there were any conflicts of interests listed that could indicate the possibility of biases. To achieve this, the researchers replaced the two guiding questions with, ''Is there a conflict of interest between the recommendations of the summary and the sponsor for any author or reviewer?,'' and the over-arching question was changed to ''Is the summary free of possible bias?''

Strengths and weaknesses

The main strength of the evaluation included the range of POC tools, including both subscription and free tools. Also, because the researchers evaluated 384 randomly selected summaries, the summaries covered a variety of topics and selection bias was kept to a minimum. As previously discussed, the evaluators interpreted questions or responses differently or looked in different parts of the summary for answers to the questions. Better communication before the study about where to look for the answers and how to uniformly define or quantify terms such as currency, amount of bias, and amount of graded evidence could have improved levels of agreement among the evaluators. Also, since the evaluators were librarians without clinical cases related to the summary topics, question 10 had to be omitted from the study.

Future research

The researchers are convinced that further evaluation of the usability of the CASE worksheet would be valuable and intend to undertake this project. While the CASE worksheet was developed due to a need for medical librarians to determine the quality of evidence summaries in POC tools, the researchers feel the tool would benefit other health sciences professionals. Input gathered from other medical librarians, clinicians, and health care students could include how helpful they would find the worksheet, what they think the priorities of categories should be on the form (including if some categories should receive more ''weight'' than others), if there are other questions worth adding, and how they would interpret the questions and response choices. Focus groups may be one medium of gathering this input about the worksheet.

Another option for future exploration would be how the CASE worksheet could support an EBM curriculum. The researchers would be particularly interested in studying if the worksheet encourages critical thinking when looking at evidence summaries. As medical librarians introduce EBM POC tools to students during the curriculum, librarians can model using the CASE worksheet to assess the quality of the treatment summaries that students find.

CONCLUSION

With the increasing number of EBM POC tools available, those who make library collection and resource decisions and those who utilize these tools to make clinical decisions may need to be more discriminating in their choices. The CASE worksheet was developed as a guide to critically analyze treatment summaries in POC tools. The researchers evaluated the reliability of the questions to create a more clear and consistent evaluative tool. Revisions to questions on the CASE worksheet were completed based on observations made through the evaluation results. Focus groups may, in the future, provide valuable input for additional improvements. However, similarly to a critical appraisal worksheet for an article, the CASE worksheet is a tool whose purpose is to appraise evidence. The researchers recognize that while reliability of the worksheet is important, the ultimate goal of using appraisal worksheets—to guide users through determining the quality and application of the evidence itself—must always remain paramount.

REFERENCES

1. Sackett DL, Rosenberg W, Gray JA, Haynes RB, Richardson WS. Evidence based medicine: what it is and what it isn't. BMJ. 1996 Jan;312(7023):71–2.
2. Critical appraisal skills programme [Internet]. [cited 28 Aug 2012]. <http://www.casp-uk.net>.
3. Bazarian JJ, Davis CO, Spillane LL, Blumstein H, Schneider SM. Teaching emergency medicine residents evidence-based critical appraisal skills: a controlled trial. Ann Emerg Med. 1999 Aug;34(2):148–54. DOI: http://dx.doi.org/10.1016/S0196-0644(99)70222-2.
4. Dorsch JL, Aiyer MK, Meyer LE. Impact of an evidence-based medicine curriculum on medical students' attitudes and skills. J Med Lib Assoc. 2004 Oct;92(4):397–406.
5. Liabsuetrakul T, Sirirak T, Boonyapipat S, Pornsawat P. Effect of continuous education for evidence-based medicine practice on knowledge, attitudes and skills of medical students. J Eval Clin Pract. 2012 Feb. DOI: http://dx.doi.org/10.1111/j.1365-2753.2012.01828.x.
6. Timm DF, Banks DE, McLarty J. Critical appraisal process: step-by-step. South Med J. 2012 Mar;105(3):144–8. DOI: http://dx.doi.org/10.1097/SMJ.0b013e31824a711f.
7. Faggion CM Jr., Tu YK. Evidence-based dentistry: a model for clinical practice. J Dent Ed. 2007;71(6):825–31.
8. Shurtz S, Foster MJ. Developing and using a rubric for evaluating evidence-based medicine point-of-care tools. J Med Lib Assoc. 2011 Jul;99(3):247–54. DOI: http://dx.doi.org/10.3163/1536-5050.99.3.012.
9. Crowe M, Sheppard L. A review of critical appraisal tools show they lack rigor: alternative tool structure is proposed. J Clin Epidemiol. 2011 Jan;64(1):79–89. DOI: http://dx.doi.org/10.1016/j.jclinepi.2010.02.008.

10. Graham ID, Calder LA, Hebert PC, Carter AO, Tetroe JM. A comparison of clinical practice guideline appraisal instruments. Int J Tech Assess Health Care. 2000 Oct;16(4):1024–38.

11. Katrak P, Bialocerkowski AE, Massy-Westropp N, Kumar S, Grimmer KA. A systematic review of the content of critical appraisal tools. BMC Med Res Methodol. 2004 Sep;4:22. DOI: http://dx.doi.org/10.1186/1471-2288-4-22.

12. Vlayen J, Aertgeerts B, Hannes K, Sermeus W, Ramaekers D. A systematic review of appraisal tools for clinical practice guidelines: multiple similarities and one common deficit. Int J Qual Health Care. 2005 Jun;17(3):235–42. DOI: http://dx.doi.org/10.1093/intqhc/mzi027.

13. Wendt O, Miller B. Quality appraisal of single-subject experimental designs: an overview and comparison of different appraisal tools. Educ Treat Child. 2012 May;35(2):235–68. DOI: http://dx.doi.org/10.1353/etc.2012.0010.

14. Cicchetti DV, Feinstein AR. High agreement but low kappa: II. resolving the paradoxes. J Clin Epidemiol. 1990;43(6):551–8. DOI: http://dx.doi.org/10.1016/0895-4356(90)90159-M.

15. Feinstein AR, Cicchetti DV. High agreement but low kappa: I. the problems of two paradoxes. J Clin Epidemiol. 1990;43(6):543–9. DOI: http://dx.doi.org/10.1016/0895-4356(90)90158-L.

16. Hollingworth W, Medina LS, Lenkinski RE, Shibata DK, Bernal B, Zurakowski D, Jarvik JG. Interrater reliability in assessing quality of diagnostic accuracy studies using the QUADAS tool: a preliminary assessment. Acad Radiol. 2006 Jul;13(7):803–10. DOI: http://dx.doi.org/10.1016/j.acra.2006.03.008.

17. Hoy D, Brooks P, Woolf A, Blyth F, March L, Bain C, Baker P, Smith E, Buchbinder R. Assessing risk of bias in prevalence studies: modification of an existing tool and evidence of interrater agreement. J Clin Epidemiol. 2012 Sep;65(9):934–9. DOI: http://dx.doi.org/10.1016/j.jclinepi.2011.11.014.

18. Kang D, Wu Y, Hu D, Hong Q, Wang J, Zhang X. Reliability and external validity of AMSTAR in assessing quality of TCM systematic reviews. Evid Based Complement Alternat Med. 2012 Feb;article ID: 732195. DOI: http://dx.doi.org/10.1155/2012/732195.

19. Shaneyfelt T, Baum KD, Bell D, Feldstein D, Houston TK, Kaatz S, Green M. Instruments for evaluating education in evidence-based practice: a systematic review. JAMA. 2006 Sep 6;296(9):1116–27. DOI: http://dx.doi.org/10.1001/jama.296.9.1116.

20. Cochrane Collaboration. Higgins JPT, Green S, eds. Cochrane handbook for systematic reviews of interventions. Chichester, England, UK: Wiley-Blackwell; 2008.

AUTHORS' AFFILIATIONS

Margaret J. Foster, MLIS, MPH, AHIP, margaretfoster@tamu.edu, Systematic Reviews and Research Coordinator; **Suzanne Shurtz, MLIS, AHIP,** sshurtz@library.tamu.edu, Instructional Services Librarian; Medical Sciences Library, Texas A&M University, 4462 TAMU, College Station, TX 77843-4462

Received August 2012; accepted December 2012